# Обучение машинного перевода без параллельных текстов*

***Ярошенко А. М. Бахтеев*[1] *О. Ю. Стрижов*[2] *В. В.***
[1]Московский физико-технический институт
[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

В данной работе исследуется задача машинного перевода между двумя языками. Для решения часто используются параллельные предложения, то есть совпадающие по смыслу фразы на двух языках. В работе рассматривается альтернативная модель, не требующая большого количества параллельных предложений. Она использует нейронную сеть типа Seq2Seq, имеющую скрытое пространство. [Тут добавится что-то от меня]. Для проверки качества модели проводится вычислительный эксперимент по переводу предложений между близкими языками, такими как русский и украинский.

**Ключевые слова**: *нейронные сети, машинный перевод, автокодировщики.*

## Введение

В зависимости от специфики пары языков выделяют несколько подходов к машинному переводу. При наличии достаточного числа параллельных предложений(порядка миллиона) использование глубоких нейронных сетей привело к получению хороших результатов [?],[?].

Но для многих пар языков нет достаточной базы примеров. Одним из подходов на основе параллельных предложений является пополнение обучающей выборки переводами с предыдущих итераций работы нейронной сети [?].

Ниже представлено решение задачи машинного перевода при отсутствии достаточного количества параллельных предложений [?], [?], [?]. В модели используются 2 типа автокодировщиков: рекуррентные нейронные сети LSTM ([?], [?]), которые реализуют перевод слов в скрытое пространство, и сеть-дискриминатор, определяющая по векторному представлению язык исходного предложения. Сети LSTM оптимизируются так, чтобы представление одного и того же предложения на разных языках совпадало в скрытом пространстве, то есть, чтобы дискриминатору было сложнее определить язык, к которому относится вектор. Обучение состоит из двух фаз. На первой оптимизируется работа дискриминатора: предложение кодируется с добавлением шума ([?]) и подаётся на вход и происходит перераспределение параметров. На второй стадии происходит перераспределение парамтеров уже у сетей-кодировщиков. После проведение этих шагов вычисляется значение функции потерь.

Такой подход был использован в [?] для пары языков французский-русский. В данной работе будет проведен схожий эксперимент для перевода с русского на украинский. Качество переводчика в работе оценивается с помощью метрики BLEU [?].

## Постановка задачи

В данной задаче в качестве обучающей выборки используются несопоставленные друг другу предложения на обоих языках $D^{src} = [s_1^{src}, ..., s_{m_{src}}^{src}]$, $D^{tgt} = [s_1^{tgt}, ..., s_{m_{tgt}}^{tgt}]$, по которым

Научный руководитель: Стрижов В. В. Авторы: А.В. Грабовой, О.Ю. Бахтеев, В.В. Стрижов, Eric Gaussier, координатор Малиновский Г. С. Консультант: Бахтеев О. Ю.

нужно предоставить перевод на другой язык. Также предоставлен блок параллельных предложений для проверки качества перевода.

Предлагается решение в виде модели из двух рекуррентых нейронных сетей для реализации декодера, дискриминатора и энкодера. Для нулевого приближения используется пословный перевод [**?**].

Ошибка модели на валидационной выборке складывается из трех составляющих: доли в целом неправильно переведнных предложений, доли ошибочно переведенных слов и accuracy, параметры которой будут подобраны в ходе эксперимента(я плохо поняла, мы выбарем одну из них? если да, то я за первую).

Так как у нас нет достаточно большого корпуса из параллельных предложений, мы будем использовать следующую схему построения модели. Первая нейронная сеть энкодер будет переводить исходное предложение в скрытое пространство, где дискриминатор будет по вектору определять, какому языку он принадлежал и соответственно использовать декодер, соответствующий другому языку. Идея этого решения в том, чтобы приблизить друг к другу пространства, соответствующие разным исходным языкам.

Для реализации этого метода определим функционалы, которые будут минимизироваться. Во-первых необходимо зашумить исходные предложения, чтобы модель не обучилась возвращать в конце цикла исходные данные. Пусть $\sigma(x)$ - результат наложения шума на слово x. На этом шаге оптимизации будет минимизироваться следующая функция:

$$L_{AE} = ||d(e(\sigma(x))) - x||^2$$

Далее на этапе использования пословного перевода функция потерь будет иметь вид:

$$L_{TR} = ||d(e(\hat{g}(e(x))) - x||^2$$

И последний этап - оптимизация дискримантора, чтобы он различал представления векторов разных языков в скрытом пространстве:

$$L_{ADV} = \log p(lang = src|Encoder(x)) + \log p(lang = tgt|Encoder(y))$$

Таким образом, имеем задачу оптимизации:

$$L = a * L_{AE} + b * L_{TR} + c * LADV \longrightarrow min$$

где a,b,c калибруемые гиперпараметры.